EVOLUTION ON DISTRIBUTIVE LATTICES

NIKO BEERENWINKEL*, NICHOLAS ERIKSSON, AND BERND STURMFELS
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720, USA
{NIKO,ERIKSSON,BERND}@MATH.BERKELEY.EDU
*CORRESPONDING AUTHOR:
PHONE: +1 (510) 642-3529, FAX: +1 (510) 642-8204

Abstract. We consider the directed evolution of a population after an intervention that has significantly altered the underlying fitness landscape. We model the space of genotypes as a distributive lattice; the fitness landscape is a real-valued function on that lattice. The risk of escape from intervention, i.e., the probability that the population develops an escape mutant before extinction, is encoded in the risk polynomial. Tools from algebraic combinatorics are applied to compute the risk polynomial in terms of the fitness landscape. In an application to the development of drug resistance in HIV, we study the risk of viral escape from treatment with the protease inhibitors ritonavir and indinavir.

Keywords: fitness landscape, distributive lattice, directed evolution, risk polynomial, chain polynomial, HIV drug resistance, Bayesian network, mutagenetic tree

1. Introduction

The evolutionary fate of a population is determined by the replication dynamics of the ensemble and by the reproductive success of its individuals. We are interested in scenarios where most individuals have a low fitness, eventually leading to extinction, and only a few types of individuals ("escape mutants") can survive permanently. These situations often arise due to a significant change of the underlying fitness landscape. For example, a virus that has been transmitted to a new host is confronted with a new immune response. Likewise, medical interventions such as radiation therapy, vaccination, or chemotherapy result in altered fitness landscapes for the targeted agents, which may be bacteria, viruses, or cancer cells.

Given a population and such a hostile fitness landscape, the central question is whether the population will survive. In the case of medical interventions we wish to know the probability of successful treatment. Answering this question involves computing the risk of evolutionary escape, i.e., the probability that the population develops an escape mutant before extinction. We present a mathematical framework for computing such probabilities.

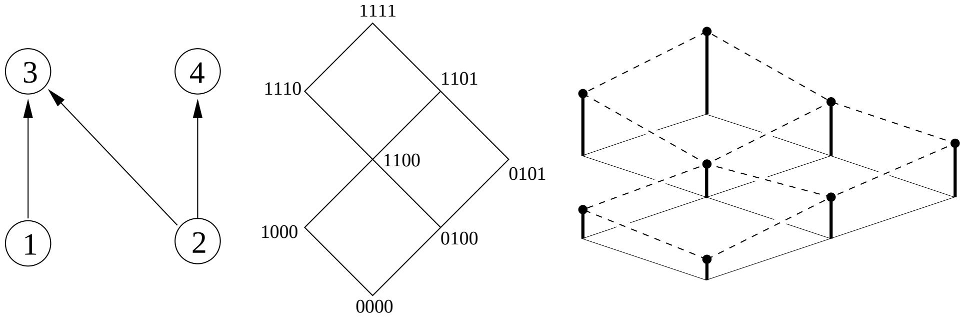

FIGURE 1. An event poset, its genotype lattice, and a fitness landscape.

Our primary application is the evolution of drug resistance during treatment of HIV infected patients [9]. We consider therapy with two different protease inhibitors (PIs). These compounds interfere with HIV particle maturation by inhibiting the viral protease enzyme. The effectiveness of PI therapy is limited by the development of drug resistance. Rapid and highly error prone replication of a large virus population generates mutants that resist the selective pressure of drug therapy. PI resistance is caused by mutations in the protease gene that reduce the binding affinity of the drug to the enzyme. These mutations have been shown to accumulate in a stepwise manner [6]. For most PIs, no single mutation confers a significant level of resistance, but multiple mutations are required for escape from drug pressure. Quantitative predictions of the probability of successful PI treatment would help in finding effective antiretroviral combination therapies. Selecting a drug combination amounts to controlling the viral fitness landscape.

We regard the directed evolution of a population towards an escape state as a fluctuation on a fitness landscape. The space of genotypes is modeled as follows. We start with a finite partially ordered set (poset) $\mathcal{E}$ whose elements are called *events*. The events are non-reversible mutations with some constraints on their order of occurrence. Such constraints are primarily due to epistatic effects between different loci in a genome [7]. The event constraints define the poset structure: $e_1 < e_2$ in $\mathcal{E}$ means that event e_1 must occur before event e_2 can occur. Each genotype g is represented by a subset of $\mathcal{E}$, namely, the set of all events that occurred to create g. Thus a genotype g is an *order ideal* in the poset $\mathcal{E}$. The space of genotypes $\mathcal{G}$ is the set of all order ideals in $\mathcal{E}$, which is a *distributive lattice* [25, Sec. 3.4]. The order relation on $\mathcal{G}$ is set inclusion and corresponds to the accumulation of mutations. This mathematical formulation is reasonable in the above situations, where a population is exposed to strong selective pressure.

The risk of escape is governed by the structure of $\mathcal{G}$, the fitness function on $\mathcal{G}$, and the population dynamics (such as the mutation rates and population size). Our focus is on the dependency of the risk of escape on the assigned fitness values for each genotype $g \in \mathcal{G}$. This leads us to the *risk* *polynomial*, which is shown to be equivalent to a well-known object in algebraic combinatorics. Indeed, one of the objectives of this work is to provide a bridge between algebraic combinatorics and evolutionary biology.

This paper is organized as follows. In Section 2 we formalize our model of a static fitness landscape on the genotype lattice $\mathcal{G}$ derived from an event poset $\mathcal{E}$, and we discuss evolution on the lattice $\mathcal{G}$. In Section 3 we review the multistate branching process studied by Iwasa, Michor and Nowak [14, 15].

In Section 4 we study the Bayesian networks which arise from identifying the events in $\mathcal{E}$ with binary random variables. These statistical models can be used to infer the genotype space from given data. For conjunctive Bayesian networks we recover the distributive lattice of order ideals in $\mathcal{E}$. Of particular interest is the case where $\mathcal{E}$ is a directed forest: here the Bayesian network is a mutagenetic tree model [3, 4]. The application of our methods to the development of PI resistance in HIV is presented in Section 5.

The Appendix summarizes various representations of the risk polynomial in terms of structures from algebraic combinatorics. Efficient methods for computing the risk polynomial and their implementation are presented.

2. Fitness landscapes on distributive lattices

A partially ordered set (or poset) is a set $\mathcal{E}$ together with a binary relation, denoted "$\leq$", which is reflexive, antisymmetric, and transitive. Here we fix a finite poset $\mathcal{E}$ whose elements are called *events*. If the number of events is n then we often identify the set underlying $\mathcal{E}$ with the set $[n] = \{1, 2, \ldots, n\}$. In this way, the subsets of $\mathcal{E}$ are encoded by the 2^n binary strings of length n. The empty subset of $\mathcal{E}$ is encoded by the all-zero string $\hat{0} = 00\cdots0$ which represents the *wild type*, and the full set $\mathcal{E}$ is encoded by the all-one string $\hat{1} = 11\cdots1$ which represents the *escape state*.

An order ideal g in a poset $\mathcal{E}$ is a subset of $\mathcal{E}$ that is closed downward; that is, if $e_2 \in g$ and $e_1 \leq e_2$, then $e_1 \in g$. The set of all order ideals of $\mathcal{E}$ forms a distributive lattice $J(\mathcal{E})$ under inclusion. Birkhoff's Representation Theorem [25, Thm. 3.4.1] states that all distributive lattices have the form $J(\mathcal{E})$ for a poset $\mathcal{E}$. We write $\mathcal{G} = J(\mathcal{E})$, and we call $\mathcal{G}$ the *genotype lattice*.

Example 1. Let $\mathcal{E}$ be the trivial poset, where no two events are comparable, with $|\mathcal{E}| = n$. Then $\mathcal{G} = J(\mathcal{E})$ is the Boolean lattice consisting of all subsets of $\mathcal{E}$ ordered by inclusion. This means that all possible combinations of mutations are possible, and they can occur in any order. Each of the 2^n binary strings $g \in \{0,1\}^n$ represents a mutational pattern, or genotype.

In general, the event poset $\mathcal{E}$ does have non-trivial relations $e_1 < e_2$. The relation $e_1 < e_2$ excludes all genotypes g with $g_{e_1} = 0$ and $g_{e_2} = 1$ from $\mathcal{G}$. The remaining genotypes g form a sublattice of the Boolean lattice $\{0,1\}^n$, and this is precisely our distributive lattice $\mathcal{G} = J(\mathcal{E})$. Note that the lattice $\mathcal{G}$ is ranked, with the rank function given by $\mathrm{rank}(g) = |g|$.

Example 2. Consider a scenario with $n = 4$ mutation events, labeled $\mathcal{E} = \{1, 2, 3, 4\}$. Suppose that event 3 can only occur after events 1 and 2, and event 4 can only occur after event 2. This allows for precisely eight genotypes

$$\mathcal{G} \;=\; \big\{0000, 1000, 0100, 1100, 0101, 1110, 1101, 1111\big\}.$$

The event poset $\mathcal{E}$ and the genotype lattice $\mathcal{G}$ are shown in Figure 1.

A fitness landscape associates to each possible genotype a number which quantifies the reproductive capacity of an individual with that genotype [21]. We define a *fitness landscape* on the distributive lattice $\mathcal{G}$ to be any function $\mathbf{f}\colon \mathcal{G} \to \mathbb{R}$. The value $\mathbf{f}(g)$ at any $g \in \mathcal{G}$ is the *fitness* of the genotype g. Thus, the space of all fitness landscapes is the finite-dimensional vector space $\mathbb{R}^{\mathcal{G}}$.

We shall consider certain special models of fitness landscapes, which are represented by linear subspaces of $\mathbb{R}^{\mathcal{G}}$. In the following definitions, a genotype g is regarded as a subset of the event poset $\mathcal{E}$, where $|\mathcal{E}| = n$. A *constant fitness landscape* has the form $\mathbf{f}(g) \equiv a$ for some constant a. Thus the constant landscapes form a line through the origin in $\mathbb{R}^{\mathcal{G}}$. A *graded fitness landscape* is a landscape on $\mathcal{G}$ whose fitness values depend only on the rank. Equivalently, we have $\mathbf{f}(g) = a_{|g|}$ for constants $a_0, a_1, \ldots, a_n$. Thus, graded fitness landscapes form an $(n+1)$-dimensional linear subspace of $\mathbb{R}^{\mathcal{G}}$.

Our biological application in Section 5 uses the graded fitness landscape model, which means that the fitness of a virus type depends only on the number of mutations it harbors. We shall model situations where a virus escapes from a wild type $\hat{0}$ to a drug-resistant type $\hat{1}$. In this case, we assume a graded fitness landscape that is monotonically increasing with rank, i.e.,

$$a_0 \;<\; a_1 \;<\; a_2 \;<\; \cdots \;<\; a_n.$$

This implies that the fitness landscape $\mathbf{f}$ has a unique local (and global) maximum at the drug resistant type $\hat{1}$, which is the top element in $\mathcal{G}$.

We next introduce the mathematical framework for evolution on a fitness landscape. The general setup is as in the work of Reidys and Stadler [21], but this is adapted here to our specific situation, where the genotypes form a distributive lattice $\mathcal{G}$. The order relation on $\mathcal{G}$, which comes from inclusion of subsets of $\mathcal{E}$, induces a neighborhood structure on $\mathcal{G}$ where the neighbors of $g \in \mathcal{G}$ are the genotypes that strictly contain g,

$$N(g) \;:=\; \big\{h \in \mathcal{G} \;\mid\; g \subset h\big\}. \tag{1}$$

Unlike the typical situation considered in [21], this notion of neighborhood is not symmetric. To be precise, we have that $h \in N(g)$ implies $g \notin N(h)$.

This neighborhood structure implies that mutational changes are possible only upward in the genotype lattice. This structure models a directed evolutionary process from the wild type $\hat{0}$ towards the escape state $\hat{1}$. Typically, our configuration space $\mathcal{G}$ is a small subset of the Boolean lattice $\{0, 1\}^n$ of all binary strings. Indeed, in the course of viral evolution, a population will visit only a small fraction of $\{0, 1\}^n$, as most mutants are not viable.

Suppose that the number of genotypes in $\mathcal{G}$ is m. We wish to define dynamics between the states of $\mathcal{G}$. To this end, we fix a linear extension of $\mathcal{G}$, and we introduce an $m \times m$ matrix of transition rates, written $\mathbf{U} = (u_{gh})$, whose rows and columns are indexed by genotypes $g, h \in \mathcal{G}$. Each entry u_{gh} of the matrix $\mathbf{U}$ is a non-negative real number which is zero unless $h \in N(g)$. In the framework of algebraic combinatorics, it is convenient to think of the matrix $\mathbf{U}$ as an element in the incidence algebra of $\mathcal{G}$; see [25, Sec. 3.6].

We further assume that the non-zero mutation rates u_{gh} depend only on the events in $h \backslash g$. Equivalently, the rate at which a collection of mutation events occurs is independent of which other mutations have already occurred. With this assumption, there are only n free parameters $\mu_1, \ldots, \mu_n$ in the matrix $\mathbf{U}$, where μ_e is the mutation rate of event e. Then

$$u_{gh} \;=\; \begin{cases} \prod_{e \in h\backslash g} \mu_e & \text{if } g \subset h \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In particular, if all rates are the same, say $\mu = \mu_1 = \cdots = \mu_n$, then the entries of $\mathbf{U}$ are $u_{gh} = \mu^{|h\backslash g|}$ if $g \subset h$ and $u_{gh} = 0$ otherwise.

Example 3. For the genotype lattice $\mathcal{G}$ in Figure 1, the matrix $\mathbf{U}$ equals

$$\begin{array}{c|cccccccc}
 & 0000 & 1000 & 0100 & 1100 & 0101 & 1110 & 1101 & 1111 \\
\hline
0000 & 0 & \mu_1 & \mu_2 & \mu_1\mu_2 & \mu_2\mu_4 & \mu_1\mu_2\mu_3 & \mu_1\mu_2\mu_4 & \mu_1\mu_2\mu_3\mu_4 \\
1000 & 0 & 0 & 0 & \mu_2 & 0 & \mu_2\mu_3 & \mu_2\mu_4 & \mu_2\mu_3\mu_4 \\
0100 & 0 & 0 & 0 & \mu_1 & \mu_4 & \mu_1\mu_3 & \mu_1\mu_4 & \mu_1\mu_3\mu_4 \\
1100 & 0 & 0 & 0 & 0 & 0 & \mu_3 & \mu_4 & \mu_3\mu_4 \\
0101 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 & \mu_1\mu_3 \\
1110 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_4 \\
1101 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_3 \\
1111 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Note that the entry in row g and column h of any power $\mathbf{U}^k$ equals u_{gh} times the number of paths of length k from g to h in $\mathcal{G}$. In particular, $\mathbf{U}^5 = 0$.

Let $\mathbf{f}$ be a fitness landscape on $\mathcal{G}$ and $\mathbf{F} = \operatorname{diag}\big(\mathbf{f}(g) \mid g \in \mathcal{G}\big)$ the $m \times m$ diagonal matrix whose entries are the fitness values. The entry of the matrix product $\mathbf{UF}$ in row g and column h represents the probability of genotype g transitioning into genotype h in one step. A precise probabilistic derivation and interpretation will be given in the next section.

We are interested in *all* mutational pathways that lead from the wild type $\hat{0}$ to the escape state $\hat{1}$. Towards this end, note that the entry (g, h) of the matrix $(\mathbf{UF})^k$ represents the probability of genotype g evolving to genotype h along any mutational pathway (chain) of length k in the genotype lattice $\mathcal{G}$. The chains from $\hat{0}$ to $\hat{1}$ in $\mathcal{G}$ are accounted for by the upper right hand entry of $(\mathbf{UF})^k$. Note that the matrix $(\mathbf{UF})^k$ is zero for $k > n$.

To account for chains of arbitrary length, we consider the matrix

$$(\mathbf{I} - \mathbf{UF})^{-1} - \mathbf{I} \;=\; \mathbf{UF} + (\mathbf{UF})^2 + (\mathbf{UF})^3 + \cdots + (\mathbf{UF})^n, \tag{3}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. We summarize our discussion in the following proposition, which is proved by elementary matrix algebra.

Proposition 4. *The entry of the matrix (3) in row g and column h is zero unless $g \subset h$, in which case it is $u_{gh} \cdot \mathbf{f}(h) \cdot P_{gh}(\mathbf{f})$ where P_{gh} is a polynomial function of degree $|h \backslash g| - 1$ on the space of all fitness landscapes $\mathbb{R}^{\mathcal{G}}$.*

The polynomial $P_{gh}(\mathbf{f})$ is the generating function for all chains from g to h in $\mathcal{G}$. This will be made precise in the following corollary. We shall restrict ourselves to the most important case when $g = \hat{0}$ is the wild type and $h = \hat{1}$ is the escape state. Studying $P_{\hat{0}\hat{1}}(\mathbf{f})$ only is no loss of generality because any interval of a distributive lattice is again a distributive lattice.

Proposition 4 tells us that $P_{\hat{0}\hat{1}}(\mathbf{f})$ is a polynomial of degree $n-1$ in the unknown fitness values $\mathbf{f}(g)$, which are also written as f_g, where $g \in \mathcal{G}$.

Corollary 5. *The polynomial $P_{\hat{0}\hat{1}}(\mathbf{f})$ in the upper-right entry of (3) equals*

$$P_{\hat{0}\hat{1}}(\mathbf{f}) \quad = \sum_{\hat{0}=g_0 \subset g_1 \subset \cdots \subset g_k = \hat{1}} f_{g_1} f_{g_2} \cdots f_{g_{k-1}}, \tag{4}$$

where the sum runs over all chains from $\hat{0}$ to $\hat{1}$ in the genotype lattice $\mathcal{G}$.

3. The risk of escape

For a poset of events $\mathcal{E}$ and the corresponding distributive lattice $\mathcal{G} = J(\mathcal{E})$, the *risk polynomial* of $\mathcal{G}$ is defined as the polynomial (4), which we denote by $\mathcal{R}(\mathcal{G}; \mathbf{f})$. The risk polynomial was introduced in [14, 15]. In this section we review the evolutionary dynamics model proposed in these papers, and we discuss the probabilistic meaning of the risk polynomial.

Example 6. Let $\mathcal{G}$ be the genotype lattice in Figure 1. Then the risk polynomial $\mathcal{R}(\mathcal{G}; \mathbf{f})$ is the following polynomial of degree three in six unknowns:

$$\begin{gathered} 1 + f_{1000} + f_{0100} + f_{1100} + f_{0101} + f_{1110} + f_{1101} \\ + f_{1000} f_{1100} + f_{0100} f_{1100} + f_{0100} f_{0101} + f_{1000} f_{1110} + f_{0100} f_{1110} \\ + f_{1000} f_{1101} + f_{0100} f_{1101} + f_{1100} f_{1110} + f_{1100} f_{1101} + f_{0101} f_{1101} \\ + f_{1000} f_{1100} f_{1110} + f_{0100} f_{1100} f_{1110} + f_{1000} f_{1100} f_{1101} \\ + f_{0100} f_{1100} f_{1101} + f_{0100} f_{0101} f_{1101}. \end{gathered}$$

If we restrict the fitness landscape $\mathbf{f}$ to lie in a linear subspace of $\mathbb{R}^{\mathcal{G}}$, then $\mathcal{R}(\mathcal{G}; \mathbf{f})$ specializes to a polynomial in fewer unknowns. For example, the risk polynomial for graded fitness landscapes is obtained from the specialization $\mathbf{f}(g) = a_{|g|}$. That risk polynomial has degree $n-1$ and is denoted by $\mathcal{R}(\mathcal{G}; a_1, \ldots, a_{n-1})$. For instance, $\mathcal{R}(\mathcal{G}; \mathbf{f})$ in Example 6 specializes to

$$\mathcal{R}(\mathcal{G}; a_1, a_2, a_3) = 1 + 2a_1 + 2a_2 + 2a_3 + 3a_1 a_2 + 4a_1 a_3 + 3a_2 a_3 + 5a_1 a_2 a_3.$$

For constant fitness landscapes $\mathbf{f} \equiv a$, the risk polynomial is a polynomial in one unknown a. It is denoted $\mathcal{R}(\mathcal{G}; a)$. In our running example,

$$\mathcal{R}(\mathcal{G}; a) \;=\; 1 + 6a + 10a^2 + 5a^3.$$

We now make precise the notion of *risk of escape*, which will justify our definition of the risk polynomial. Our derivation is based on the model for the dynamics of a replicating population on a fitness landscape studied by Iwasa, Michor and Nowak [14, 15]. See also the work of Wilke [29] and the references given therein for approaches to computing fixation probabilities.

A *multistate branching process* [1] consists of a set of genotypes along with a fitness landscape and mutation rates between genotypes. We assume a discrete time process, where in one generation an individual with genotype g has a random number of offspring following a Poisson distribution with mean R_g. Some of these offspring may be mutants according to the mutation rates u_{gh}. The parameter R_g is the *basic reproductive ratio* [19, Chap. 3].

We assume there is no interaction between individuals; each reproduces at a rate independent of the distribution of the population. Let $\rho^k_{g,h}$ be the probability that one individual of genotype g has k children of type h. Then,

$$\rho^k_{g,h} \;=\; \frac{(u_{gh}R_g)^k \cdot e^{-u_{gh}R_g}}{k!}. \tag{5}$$

The *reproductive fitness* f_g is related to the reproductive ratio R_g by

$$f_g \;=\; \frac{R_g}{1 - R_g} \qquad \text{and} \qquad R_g \;=\; \frac{f_g}{1 + f_g}. \tag{6}$$

Let ξ_g be the probability of escape starting with one individual of genotype g, so $1-\xi_g$ is the probability of extinction. In particular, $\xi_{\hat{1}}$ is the probability that one resistant virus will not become extinct. Each of these probabilities is a function of the mutation rates u_{gh} and the reproductive ratios R_g. We assume that the u_{gh} are as in (2), but with $u_{gg} = 1$. Thus, each escape probability ξ_g can be expressed as a function of the μ_e for $e \in \mathcal{E}$ and (using the relation (6)) the fitness values f_g for $g \in \mathcal{G}$.

Theorem 7. *If $\xi_g \ll 1$ for $g \neq \hat{1}$, then the probability of escape on the fitness landscape $\mathbf{f} \in \mathbb{R}^{\mathcal{G}}$ starting with one individual of wild type $\hat{0}$, satisfies*

$$\xi_{\hat{0}} \quad\approx\quad \xi_{\hat{1}} \cdot f_{\hat{0}} \cdot \prod_{e\in\mathcal{E}} \mu_e \cdot \mathcal{R}(\mathcal{G}; \mathbf{f}). \tag{7}$$

Proof. The probability of extinction satisfies the recursive formula

$$1 - \xi_g \quad=\quad \prod_{h \supseteq g} \sum_{k=0}^{\infty} (1 - \xi_h)^k \cdot \rho^k_{g,h}. \tag{8}$$

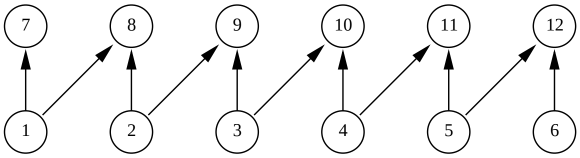

FIGURE 2. Example of an event poset whose general risk polynomial is of degree 11 in 375 unknowns.

Using (5), the right hand side of (8) can be rewritten as follows:

$$\prod_{h \supseteq g} \exp((1-\xi_h)u_{gh}R_g) \cdot \exp(-u_{gh}R_g) \quad = \quad \exp\left(\sum_{h \supseteq g} -\xi_h u_{gh} R_g\right).$$

We conclude that

$$\log(1-\xi_g) \quad = \quad -\sum_{h \supseteq g} \xi_h u_{gh} R_g \qquad \text{for all } g \in \mathcal{G}.$$

Under the assumption that $\xi_g \ll 1$ for $g \neq \hat{1}$, we can linearize the logarithms using the relation $\log(1-\xi_g) \approx -\xi_g$. This implies, for $g \in \mathcal{G}\backslash\{\hat{1}\}$,

$$\begin{aligned} \xi_g \quad & \approx \quad R_g \cdot \textstyle\sum_{h \supseteq g} \xi_h u_{gh} \\ & = \quad \tfrac{R_g}{1-R_g u_{gg}} \cdot \textstyle\sum_{h \supset g} \xi_h u_{gh} \\ & = \quad f_g \cdot \textstyle\sum_{h \supset g} \xi_h u_{gh}. \end{aligned}$$

The theorem now follows by setting $g = \hat{0}$ and expanding the last equation recursively. Here we are using the fact from (2) that the product of the u_{gh} over any chain from $\hat{0}$ to $\hat{1}$ in $\mathcal{G}$ equals $\prod_{e \in \mathcal{E}} \mu_e$. □

The typical situation of interest is a fitness landscape for which only the escape state has a basic reproductive ratio greater than one, i.e.,

$$R_{\hat{1}} > 1 \qquad \text{and} \qquad R_g < 1 \quad \text{for all} \quad g \neq \hat{1}.$$

When the positive numbers R_g are very small for $g \in \mathcal{G}\backslash\{\hat{1}\}$ then the approximation (7) is valid, and it shows the crucial role that the risk polynomial $\mathcal{R}(\mathcal{G};\mathbf{f})$ plays in assessing the risk of escape from the wild type $\hat{0}$ to the escape state $\hat{1}$. The theorem implies that the risk of escape of a population of N wild type viruses is $(1-\xi_{\hat{0}})^N$. In Section 6 we discuss the situation in which the population is not homogeneous at the time of intervention.

The risk of escape is an important quantity in analyzing the invasiveness of pathogens and in assessing the success probability of medical interventions such as chemotherapy. However, putting this concept into practice depends on our ability to actually compute the risk polynomial. It turns out that methods from algebraic combinatorics lead to efficient algorithms for this task. In the Appendix, several methods are presented in detail.

Our method of choice from a practical perspective relies on computing linear extensions of the event poset $\mathcal{E}$ (Theorem 15, Appendix). Our software implementation is available at `http://bio.math.berkeley.edu/riskpoly/` . For an example of the efficiency of the software, let $\mathcal{E}$ be the poset in Figure 2 on $n = 12$ events with cover relations $i < 6 + i$ for $1 \leq i \leq 6$ and $i < 7+i$ for $1 \leq i \leq 5$. Here the genotype lattice $\mathcal{G}$ consists of 375 genotypes. The risk polynomial $\mathcal{R}(\mathcal{G}; \mathbf{f})$ is a polynomial of degree 11 in 375 unknowns f_g. This polynomial has 224,750,298 monomials in the 375 unknowns, but we represent it as a sum of 2,702,765 products, one for each linear extension of the event poset $\mathcal{E}$. Our software takes about ten seconds to compute this representation of $\mathcal{R}(\mathcal{G}; \mathbf{f})$. The result takes up 200MB of disk space.

The univariate risk polynomial for this example is

$$1 + 375a + 19088a^2 + 324498a^3 + 2610169a^4 + 11729394a^5 + 32080336a^6 + \\ 55597909a^7 + 61448965a^8 + 42020208a^9 + 16216590a^{10} + 2702765a^{11}.$$

Thus, exact symbolic computations, as opposed to numerical approximations, may be necessary and feasible when one is interested in assessing the risk of escape in applications like the one described in Section 5 below.

4. Distributive lattices from Bayesian networks

In this section, we present a family of statistical models that naturally gives rise to distributive lattices. This statistical interpretation provides a method for deriving the genotype lattice $\mathcal{G}$ directly from data. The basic idea is to estimate the poset structure on $\mathcal{E}$ from observed genotypes, by applying model selection techniques to a range of Bayesian networks, and to define $\mathcal{G}$ as the set of all genotypes with non-zero probability in the model.

We first make precise the derivation of a genotype space from a statistical model. Let $\mathcal{E}$ be an unordered set of n genetic events. The events are labeled by $1, 2, \ldots, n$. Subsets of $\mathcal{E}$ are identified with binary strings $g \in \{0, 1\}^n$. They are the possible genotypes. We consider binary random variables $X_{\mathcal{E}} = (X_1, \ldots, X_n)$, where $X_e = 1$ indicates the occurrence of event e. Let Δ denote the $(2^n - 1)$-dimensional simplex of probability distributions on $\{0, 1\}^n$. A *statistical model* for $X_{\mathcal{E}}$ is a map $p \colon \Theta \to \Delta$, where Θ is some parameter space. The g-th coordinate of p, denoted p_g, is the probability of genotype $g \in \{0, 1\}^n$ under the model p. The *induced genotype space* of the model $p \colon \Theta \to \Delta$ is the set $\mathcal{G}_p$ of all strings $g \in \{0, 1\}^n$ such that p_g is not the zero function on Θ. We regard $\mathcal{G}_p$ as a poset ordered by inclusion.

Now consider a directed acyclic graph on the set of events $\mathcal{E}$. We will also call this graph $\mathcal{E}$. The *Bayesian network model*, or directed acyclic graphical model, defined by $\mathcal{E}$ is the family of joint distributions that factor as

$$\Pr(X_1, \ldots, X_n) \quad = \quad \prod_{e \in \mathcal{E}} \Pr(X_e \mid X_{\mathrm{pa}(e)}),$$

where $\mathrm{pa}(e)$ denotes the set of parents of e in $\mathcal{E}$. Equivalently, a Bayesian network is specified by a set of conditional independence statements. Each node is independent of its ancestors given its parents. See [16] for an introduction to the relevant statistical theory and [13] for an algebraic perspective.

The parameters for a Bayesian network are specified by providing, for each event $e \in \mathcal{E}$, a $2^{|\mathrm{pa}(e)|} \times 2$ matrix θ^e. The matrix entries are

$$\theta^e_{g_{\mathrm{pa}(e)},g_e} \quad = \quad \Pr\left(X_e = g_e \mid X_{\mathrm{pa}(e)} = g_{\mathrm{pa}(e)}\right),$$

for $g_{\mathrm{pa}(e)} \in \{0,1\}^{\mathrm{pa}(e)}$, $g_e \in \{0,1\}$. These conditional probabilities satisfy

$$\theta^e_{g_{\mathrm{pa}(e)},0} \geq 0\,, \;\; \theta^e_{g_{\mathrm{pa}(e)},1} \geq 0 \quad \text{and} \quad \theta^e_{g_{\mathrm{pa}(e)},0} + \theta^e_{g_{\mathrm{pa}(e)},1} \; = \; 1. \tag{9}$$

Set $d = \sum_{e \in \mathcal{E}} 2^{|\mathrm{pa}(e)|}$ and $\Theta = [0,1]^d$. The points in the cube Θ are identified with n-tuples of matrices $\theta = (\theta^e \,|\, e \in \mathcal{E})$ as above. The *general Bayesian network* is the polynomial map $p\colon \Theta \to \Delta$ whose coordinates are

$$p_g(\theta) \;\; = \;\; \prod_{e \in \mathcal{E}} \theta^e_{g_{\mathrm{pa}(e)},g_e}. \tag{10}$$

The general Bayesian network on $\mathcal{E}$ induces the genotype space $\mathcal{G}_p = \{0,1\}^n$, the Boolean lattice on $\mathcal{E}$. Indeed, the factorization (10) implies that no genotype $g \in \{0,1\}^n$ has probability zero for all parameter values.

To obtain other genotype spaces, we replace the cube $\Theta = [0,1]^d$ by one of its faces, as follows. For each event $e \in \mathcal{E}$ consider a Boolean function $\beta_e\colon \{0,1\}^{\mathrm{pa}(e)} \to \{0,1\}$. If $\beta_e(g_e) = 0$ then the row of the $2^{|\mathrm{pa}(e)|} \times 2$-matrix θ^e indexed by the genotype g is fixed to be the vector $(1,0)$; otherwise that row remains indeterminate subject to the constraints (9). Let Θ^β denote the face of Θ determined by these requirements and $p^\beta\colon \Theta^\beta \to \Delta$ the restriction of the polynomial map p to Θ^β. The resulting model is the Bayesian network on $\mathcal{E}$ constrained by the Boolean functions β^e.

If all Boolean functions β^e are disjunctions then we get the *disjunctive Bayesian network* on $\mathcal{E}$. In this model, an event e can only occur if at least one of its parent events has already occurred. If all Boolean functions β^e are conjunctions then we get the *conjunctive Bayesian network* on $\mathcal{E}$. In this model, an event e can only occur if all of its parent events have already occurred. These restricted Bayesian network models induce interesting genotype spaces. Our main result in this section concerns the conjunctive case.

We regard the given directed acyclic graph $\mathcal{E}$ as a poset by setting $e_1 \leq e_2$ if there exists a path from e_1 to e_2. We write $p^{\mathrm{conj}}\colon [0,1]^n \to \Delta$ for the conjunctive Bayesian network on $\mathcal{E}$, since it has precisely n free parameters.

Theorem 8. *The genotype space induced by the conjunctive Bayesian network on $\mathcal{E}$ is the distributive lattice of order ideals in $\mathcal{E}$, i.e., $\mathcal{G}_{p^{\mathrm{conj}}} = J(\mathcal{E})$.*

Proof. The possible genotypes g are binary strings whose coordinates g_e indicate whether or not the event e has occurred. If p is any of the Bayesian network models discussed above, then (10) implies that $g \in \mathcal{G}_p$ if and only if each $\theta^e_{g_{\mathrm{pa}(e)},g_e}$ is non-zero. Consider now the conjunctive model $p = p^{\mathrm{conj}}$. Here, the conditional probability $\theta^e_{g_{\mathrm{pa}(e)},g_e}$ is non-zero if and only if $g_e = 1$ implies $g_{\mathrm{pa}(e)} = (1,\ldots,1)$. This is precisely the condition for g to be an order ideal in $\mathcal{E}$. Thus $\mathcal{G}_p$ is the distributive lattice of order ideals of $\mathcal{E}$. □

The following example illustrates Theorem 8, and it compares the genotype spaces induced by the disjunctive and the conjunctive Bayesian network. The former is not a distributive lattice, but the latter always is.

Example 9. Let $\mathcal{E}$ be the event poset in Figure 1. The general Bayesian network model defined by $\mathcal{E}$ is parametrized by the following four matrices:

$$\begin{array}{l} \theta^1 = \begin{pmatrix} a & 1-a \end{pmatrix}, \\[1ex] \theta^2 = \begin{pmatrix} b & 1-b \end{pmatrix}, \end{array} \qquad \theta^3 = \begin{pmatrix} c_{00} & 1-c_{00} \\ c_{01} & 1-c_{01} \\ c_{10} & 1-c_{10} \\ c_{11} & 1-c_{11} \end{pmatrix}, \quad \theta^4 = \begin{pmatrix} d_0 & 1-d_0 \\ d_1 & 1-d_1 \end{pmatrix}.$$

The map $p\colon [0,1]^8 \to \Delta$ has coordinates

$$\begin{aligned}
p_{0000} &= abc_{00}d_0, & p_{0001} &= abc_{00}(1-d_0),\\
p_{0010} &= ab(1-c_{00})d_0, & p_{0011} &= ab(1-c_{00})(1-d_0),\\
p_{0100} &= a(1-b)c_{01}d_1, & p_{0101} &= a(1-b)c_{01}(1-d_1),\\
p_{0110} &= a(1-b)(1-c_{01})d_1, & p_{0111} &= a(1-b)(1-c_{01})(1-d_1),\\
p_{1000} &= (1-a)bc_{10}d_0, & p_{1001} &= (1-a)bc_{10}(1-d_0),\\
p_{1010} &= (1-a)b(1-c_{10})d_0, & p_{1011} &= (1-a)b(1-c_{10})(1-d_0),\\
p_{1100} &= (1-a)(1-b)c_{11}d_1, & p_{1101} &= (1-a)(1-b)c_{11}(1-d_1),\\
p_{1110} &= (1-a)(1-b)(1-c_{11})d_1, & p_{1111} &= (1-a)(1-b)(1-c_{11})(1-d_1).
\end{aligned}$$

This model induces the Boolean lattice $\{0,1\}^4$ as genotype space.

The disjunctive Bayesian network is the six-dimensional submodel obtained by setting $c_{00} = 1$ and $d_0 = 1$. This substitution implies

$$p_{0001} = p_{0010} = p_{0011} = p_{1001} = p_{1011} = 0.$$

The genotype space $\mathcal{G}_{p^{\mathrm{disj}}}$ consists of the remaining eleven strings in $\{0,1\}^4$. Note that $\mathcal{G}_{p^{\mathrm{disj}}}$ is not a lattice because it is not closed under intersections. For instance, 1010 and 0110 are in $\mathcal{G}_{p^{\mathrm{disj}}}$ but $0010 = 1010 \cap 0110 \notin \mathcal{G}_{p^{\mathrm{disj}}}$.

The conjunctive Bayesian network is the four-dimensional submodel obtained by setting $c_{00} = c_{01} = c_{10} = d_0 = 1$. The remaining eight non-zero probabilities are indexed by the eight genotypes in Figure 1:

$$\begin{aligned}
p_{0000} &= ab\,, & p_{0100} &= a(1-b)d_1\,,\\
p_{0101} &= a(1-b)(1-d_1)\,, & p_{1000} &= (1-a)b\,,\\
p_{1100} &= (1-a)(1-b)c_{11}d_1\,, & p_{1101} &= (1-a)(1-b)c_{11}(1-d_1)\,,\\
p_{1110} &= (1-a)(1-b)(1-c_{11})d_1\,, & p_{1111} &= (1-a)(1-b)(1-c_{11})(1-d_1).
\end{aligned}$$

If $\mathcal{E}$ is a directed forest, i.e., if every $e \in \mathcal{E}$ has at most one parent, then we can augment $\mathcal{E}$ to a tree $\mathcal{E}^T$ by adding an auxiliary root node 0 which points to the roots (edges with no parents) of the forest. On the resulting tree $\mathcal{E}^T$ we consider the *mutagenetic tree model* of [4, 11].

Proposition 10. *If $\mathcal{E}$ is a directed forest then the following three statistical models coincide: the disjunctive Bayesian network on $\mathcal{E}$, the conjunctive Bayesian network on $\mathcal{E}$, and the mutagenetic tree model on $\mathcal{E}^T$.*

Proof. The disjunctive and the conjunctive networks coincide because they are defined by the same specializations of the parameters θ^e. The identification with the mutagenetic tree model follows from [3, Thm. 14.6]. □

Mutagenetic tree models can be learned from observed data by an efficient combinatorial algorithm. With appropriate edge weights that depend on the pairwise probabilities of events, a mutagenetic tree can be obtained as the maximum weight branching rooted at 0 in the complete graph on $\{0, \ldots, n\}$; see [11]. This gives an efficient method for learning the poset $\mathcal{E}$, and hence the genotype lattice $\mathcal{G} = J(\mathcal{E})$, from data. It would be interesting to extend this model selection technique to arbitrary conjunctive Bayesian networks.

5. Applications to HIV drug resistance

We investigate the development of resistance during treatment of HIV infected patients with two different PIs. Consider the seven genetic events

$$\mathcal{E} = \{\text{K20R, M36I, M46I, I54V, A71V, V82A, I84V}\},$$

where K20R stands for the amino acid change from lysine (K) to arginine (R) at position 20 of the protease chain, etc. The occurrence of these mutations confers broad cross-resistance to the entire class of PIs. Appearance of the virus with all 7 mutations renders most of the PIs ineffective for subsequent treatment. We analyze the risk of reaching this escape state under therapy with the PIs ritonavir (RTV) and indinavir (IDV) [10, 18].

We use mutagenetic trees for estimating preferred mutational pathways and for defining genotype lattices. For both drugs, a tree $\mathcal{E}^T$ is learned from genotypes derived from patients under the respective therapy. We used 112 and 691 samples from the Stanford HIV Drug Resistance Database [22] for ritonavir and indinavir, respectively. Figure 3 shows the inferred mutagenetic trees. The models indicate that the evolution of ritonavir resistance is partly a linear process, whereas indinavir resistance develops in a less ordered fashion. This is consistent with previous studies [10, 18]. The genotype lattices $\mathcal{G}$ have size 16 for ritonavir and 45 for indinavir. We study the risk polynomials on these lattices under different fitness landscape models.

For the constant fitness landscape on $\mathcal{G}\backslash\{\hat{0}, \hat{1}\}$, we obtain

$$\begin{aligned}\mathcal{R}_{\mathrm{RTV}}(a) &= 15a^6 + 70a^5 + 131a^4 + 124a^3 + 61a^2 + 14a + 1,\\ \mathcal{R}_{\mathrm{IDV}}(a) &= 420a^6 + 1470a^5 + 1970a^4 + 1250a^3 + 372a^2 + 43a + 1.\end{aligned}$$

Thus, the risk of developing all seven PI resistance mutations is higher under indinavir therapy than under ritonavir: $\mathcal{R}_{\mathrm{IDV}}(a) > \mathcal{R}_{\mathrm{RTV}}(a)$ for $a > 0$.

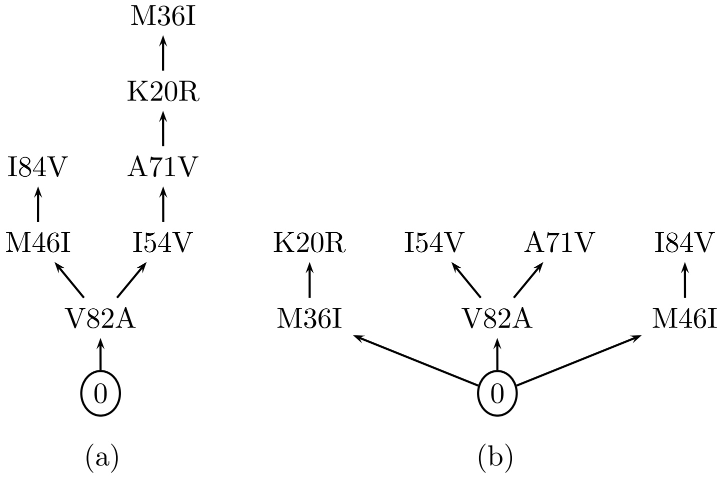

FIGURE 3. Mutagenetic tree $\mathcal{E}^T$ for the development of resistance to (a) ritonavir and (b) indinavir in the HIV-1 protease. The event poset $\mathcal{E}$ is obtained by removing the root node "0".

Intuitively, the risk under ritonavir is lower because the mutations must occur in a certain order. Likewise, the high risk under indinavir results from many mutations occurring independently, which gives rise to a large genotype lattice and to many mutational pathways from the wild type to the escape state.

More realistic fitness landscapes may be derived by modeling viral fitness as a function of drug concentration. We follow the approach pursued in [26] and use a simple saturation function for this dependency. Specifically, we assume viral fitness to be the following function of drug concentration D,

$$f_g(D) \quad = \quad \frac{\phi_g}{1 + D/r_g}, \tag{11}$$

where ϕ_g denotes the fitness of genotype g in the absence of drug and r_g the IC$_{50}$ value of g, i.e., the drug concentration necessary to inhibit viral replication *in vitro* by 50%. The IC$_{50}$ value is a measure of resistance. We will assume throughout that all $\phi_g \equiv \phi$ are equal. If we assume, in addition, that the resistance landscape is constant on $\mathcal{G}\backslash\{\hat{0}, \hat{1}\}$, with $r_g \equiv r$, then the substitution (11) turns the risk polynomial into a rational function in ϕ, D, and r. For example, for ritonavir, this rational function is

$$\frac{(15\phi^2 r^2 + 10\phi D r + 10\phi r^2 + D^2 + 2Dr + r^2)(\phi r + D + r)^4}{(D+r)^6}.$$

In general, the IC$_{50}$ values r_g are distinct and can be determined experimentally for some genotypes by phenotypic resistance testing [28], and may be predicted for all genotypes using regression techniques [2]. PI phenotypic resistance data suggests a graded resistance landscape; see [6] and [10,

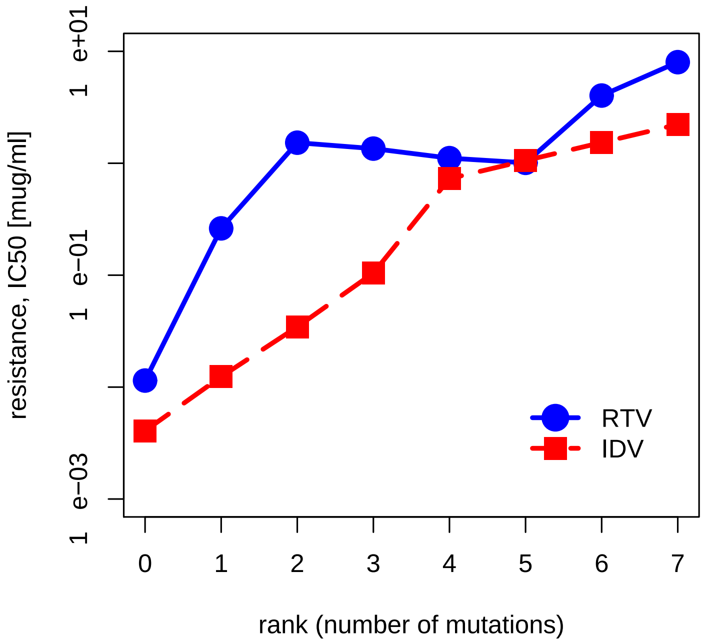

FIGURE 4. Graded resistance landscapes for ritonavir (RTV, bullets) and indinavir (IDV, squares). Resistance is quantified as the drug concentration necessary to inhibit viral replication *in vitro* by 50% (IC_{50}).

Tab. 3]. Hence, we estimate the resistance $r \in \mathbb{R}^8$ for ritonavir and indinavir by defining r_k as the mean predicted IC_{50} of all genotypes of rank k. The resulting resistance landscapes are shown in Figure 4.

The graded risk polynomials $\mathcal{R}(a_1, a_2, a_3, a_4, a_5, a_6)$ have 64 terms. After substituting $a_k = \phi/(1 + D/r_k)$, we obtain rational risk functions in D with parameter ϕ. Figure 5 illustrates the dependency of the risk on drug concentration for three different values of ϕ. For both drugs we indicate published mean plasma trough ($C_{\min}$) and peak ($C_{\max}$) levels observed in clinical settings.

This example illustrates how the risk polynomial can be used to study viral escape as a function of different parameters. For instance, given a pharmacokinetics model of antiretroviral drug therapy, we can compute the risk of developing resistance after a patient has missed a dose. Thus, our mathematical framework may help in designing robust drug combinations.

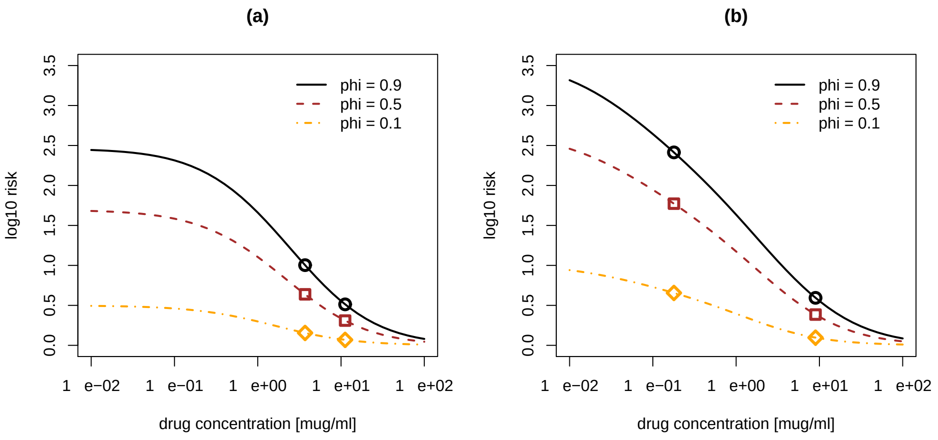

Figure 5. Drug dependent risk. The log of the risk polynomial for ritonavir (a) and indinavir (b) is displayed as a function of plasma drug concentration D. Marked values denote mean trough (C_{min}) and peak (C_{max}) levels observed in clinical studies. The parameter ϕ is the relative fitness of mutants as compared to the wild type in the absence of drug.

6. Discussion

We have presented a computational framework for assessing the risk of escape of an evolving population of pathogens. The risk of escape is the probability that the population reaches an escape state before extinction. In virus transmissions, for example, this probability is the chance of survival in the new host. In the situation of antiretroviral therapy, the risk of escape is the probability of therapy failure due to the development of drug resistance.

The general setup we consider for computing the risk of escape includes an event poset, a fitness landscape on its induced genotype lattice, and a branching process on this lattice. The event poset $\mathcal{E}$ consists of all mutational events that can occur and encodes the constraints which apply to their order of occurrence. From this structure the genotype space $\mathcal{G}$ is obtained by considering all mutational pathways that respect the order constraints. This natural construction endows $\mathcal{G}$ with the mathematical structure of a distributive lattice. The risk polynomial, the crucial factor in computing the risk of escape, turns out to coincide with the chain polynomial of the genotype lattice. We have presented methods from algebraic combinatorics that exploit this connection and that result in efficient algorithms.

The space of genotypes may also be inferred from observed genotype data using statistical model selection tools. We have identified a class of Bayesian network models, the conjunctive Bayesian networks, whose support induces a genotype lattice. Mutagenetic tree models arise as important special cases.

Here, both statistical model selection and risk computation are particularly efficient, and readily available with existing software [5] coupled with our implementation of the linear extensions method (Theorem 15, Appendix).

We have focused on the dependency of the risk polynomial on the fitness landscape and considered throughout a homogeneous wild type population prior to intervention. However, the risk of escape is calculated similarly for a quasispecies distribution at the time of intervention. In fact, this involves computing the risk polynomial of the prior fitness landscape [14]. In contrast, the branching process model can not account for recombination, horizontal gene transfer, or frequency dependent selection, since evolution is assumed to take place in multiple lineages independently.

The main challenge in using our method to compute the risk of escape from antiretroviral therapy lies in accurately modeling the fitness landscape. The dependency (11) of the fitness on drug concentration may be improved by experimentally determined viral replicative capacities in the absence of drugs. An alternative approach to derive a fitness landscape for HIV-1 proteases is based on estimating the binding affinity of the drug to the mutant protease, and the mutant's ability to cleave its natural substrates [23]. These calculations are based on simplified molecular modeling techniques. The resulting fitness landscape does not account for different drug levels, but it is independent of experimental resistance and fitness data.

Escape from indinavir and ritonavir therapy may in some cases involve mutations other than the seven we considered, although those are the most frequent mutations observed after therapy failure [10, 18]. On the other hand, viral escape might be accomplished with genotypes that harbor fewer than all of the mutations. Thus it would be desirable to compute the risk of reaching any of several escape states, rather than only the $11 \cdots 1$ type. This computation will involve similar techniques to those presented in Section 3 and the Appendix.

Finally, the PIs form only one out of four distinct classes of antiretroviral drugs that are in current clinical use. The standard of care is combination therapy with at least three different drugs from two different drug classes. Modeling the fitness landscape of combination therapy in terms of viral drug resistance and drug exposure is even more challenging, but can eventually help in designing optimal antiretroviral therapies. Algebraic combinatorics offers tools for the mathematical analysis of these biomedical problems.

Acknowledgements

Niko Beerenwinkel is supported by Deutsche Forschungsgemeinschaft under grant No. BE 3217/1-1. Nicholas Eriksson and Bernd Sturmfels are supported by the U.S. National Science Foundation, under the grants EF-0331494 and DMS-0456960 respectively, and by the DARPA program *Fundamental Laws in Biology* (HR0011-05-1-0057).

References

[1] K.B. Athreya and P.E. Ney. *Branching processes.* Dover, Mineola, New York, 1972.

[2] N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucl. Acids Res.*, 31(13):3850–3855, Jul 2003.

[3] N. Beerenwinkel and M. Drton. Mutagenetic tree models. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, chapter 14, pages 278–290. Cambridge University Press, Cambridge, UK, 2005.

[4] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.*, 12(6):584–598, 2005.

[5] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, May 2005.

[6] B. Berkhout. HIV-1 evolution under pressure of protease inhibitors: Climbing the stairs of viral fitness. *J. Biomed. Sci.*, 6:298–305, 1999.

[7] S. Bonhoeffer, C. Chappey, N.T. Parkin, J.M. Whitcomb, and C.J. Petropoulos. Evidence for Positive Epistasis in HIV-1. *Science*, 306:1547–1550, 2004.

[8] G. Brightwell and P. Winkler. Counting linear extensions. *Order*, 8(3):225–242, 1991.

[9] F. Clavel and A.J. Hance. HIV drug resistance. *N. Engl. J. Med.*, 350(10):1023–1035, Mar 2004.

[10] J.H. Condra, D.J. Holder, W.A. Schleif, O.M. Blahy, R.M. Danovich, L.J. Gabryelski, D.J. Graham, D. Laird, J.C. Quintero, A. Rhodes, H.L. Robbins, E. Roth, M. Shivaprakash, T. Yang, J.A. Chodakewitz, P.J. Deutsch, R.Y. Leavitt, F.E. Massari, J.W. Mellors, K.E. Squires, R.T. Steigbigel, H. Teppler, and E.A. Emini. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *J. Virol.*, 70(12):8270–8276, 1996.

[11] R. Desper, F. Jiang, O.P. Kallioniemi, H. Moch, C.H. Papadimitriou, and A.A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.*, 6(1):37–51, 1999.

[12] R. Ehrenborg. On posets and Hopf algebras. *Adv. Math.*, 119(1):1–25, 1996.

[13] L. Garcia, M. Stillman, and B. Sturmfels. Algebraic geometry of Bayesian networks. *J. Symbol. Comput.*, 39:331–355, 2005.

[14] Y. Iwasa, F. Michor, and M.A. Nowak. Evolutionary dynamics of escape from biomedical intervention. *Proc. Biol. Sci.*, 270(1533):2573–2578, Dec 2003.

[15] Y. Iwasa, F. Michor, and M.A. Nowak. Evolutionary dynamics of invasion and escape. *J. Theor. Biol.*, 226(2):205–214, Jan 2004.

[16] S.L. Lauritzen. *Graphical Models.* Clarendon Press, 1996.

[17] E. Miller and B. Sturmfels. *Combinatorial commutative algebra*, volume 227 of *Graduate Texts in Mathematics.* Springer, New York, 2005.

[18] A. Molla, M. Korneyeva, Q. Gao, S. Vasavanonda, P.J. Schipper, H.M. Mo, M. Markowitz, T. Chernyavskiy, P. Niu, N. Lyons, A. Hsu, G.R. Granneman, D.D. Ho, C.A. Boucher, J.M. Leonard, D.W. Norbeck, and D.J. Kempf. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat. Med.*, 2(7):760–766, Jul 1996.

[19] M.A. Nowak and R.M. May. *Virus dynamics.* Oxford University Press, 2000.

[20] G. Pruesse and F. Ruskey. Generating linear extensions fast. *SIAM J. Comput.*, 23(2):373–386, 1994.

[21] C.M. Reidys and P.F. Stadler. Combinatorial landscapes. *SIAM Review*, 44:3–54, 2002.

[22] S.-Y. Rhee, M.J. Gonzales, R. Kantor, B.J. Betts, J. Ravela, and R.W. Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucl. Acids Res.*, 31(1):298–303, Jan 2003.

[23] C.D. Rosin, R.K. Belew, G.M. Morris, A.J. Olson, and D.S. Goodsell. Coevolutionary analysis of resistance-evading peptidomimetic inhibitors of HIV-1 protease. *Proc. Natl. Acad. Sci. U. S. A.*, 96:1369–1374, 1999.

[24] R.P. Stanley. A matrix for counting paths in acyclic digraphs. *J. Combin. Theory Ser. A*, 74(1):169–172, 1996.

[25] R.P. Stanley. *Enumerative combinatorics. Vol. 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997.

[26] N.I. Stilianakis, C.A. Boucher, M.D. De Jong, R. Van Leeuwen, R. Schuurman, and R.J. De Boer. Clinical data sets of human immunodeficiency virus type 1 reverse transcriptase resistant mutants explained by a mathematical model. *J. Virol.*, 71(1):161–168, 1997.

[27] Y.L. Varol and D. Rotem. An algorithm to generate all topological sorting arrangements. *Comput. J.*, 24(1):83–84, 1981.

[28] H. Walter, B. Schmidt, K. Korn, A. M. Vandamme, T. Harrer, and K. Überla. Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *J. Clin. Virol.*, 13:71–80, 1999.

[29] C.O. Wilke. Probability of fixation of an advantageous mutant in a viral quasispecies. *Genetics*, 163:467–474, 2003.

Appendix: Mathematics and computation of the risk polynomial

Here we discuss in more detail mathematical properties of the risk polynomial and we present several methods for computing it. The given data consists of an n element poset $\mathcal{E}$ and its induced genotype lattice $\mathcal{G}$, which is the distributive lattice of order ideals in $\mathcal{E}$. We assume that $\mathcal{G}$ has m elements, which are encoded either as subsets of $\mathcal{E}$ or as binary strings in $\{0,1\}^n$. The risk polynomial is the polynomial $\mathcal{R}(\mathcal{G};\mathbf{f})$ in the m unknowns $f_g = \mathbf{f}(g)$, one for each genotype g. We are also interested in specializations of $\mathcal{R}(\mathcal{G};\mathbf{f})$ obtained by setting some (or all) of the unknowns equal to each other, such as the graded risk polynomial and the univariate risk polynomial.

Stanley's linear algebra method. A direct method for computing the risk polynomial is given in Section 3. Namely, we can set all μ_e equal to one in the matrix $\mathbf{U}$ and then compute the upper right entry of the matrix $(\mathbf{I}-\mathbf{UF})^{-1}-\mathbf{I}$ of equation (3). In practice, one would compute this entry by a dynamic program which runs in time $O(m^2)$. That dynamic program is easily derived by resolving the recursion in the last equation of the proof of Theorem 7.

The following alternative linear algebra technique for computing polynomials similar to our risk polynomials was given by Stanley in [24]. Let $\mathcal{G}' = \mathcal{G}\backslash\{\hat{0},\hat{1}\}$ denote the genotype lattice with the top element $\hat{1}$ and the bottom element $\hat{0}$ removed. We define $\mathbf{A}$ to be the *anti-adjacency matrix* of the truncated genotype lattice $\mathcal{G}'$. Thus $\mathbf{A}$ is the $(m-2)\times(m-2)$-matrix with rows and columns indexed by $\mathcal{G}'$, and whose entry in row g and column

h is 0 if $g \subset h$ and is 1 otherwise. We write $\mathbf{I}$ for the $(m-2) \times (m-2)$ identity matrix and $\mathbf{F}' = \mathrm{diag}\big(\mathbf{f}(g) \,|\, g \in \mathcal{G}'\big)$ for the $(m-2) \times (m-2)$-diagonal matrix whose entries are the fitness values. Stanley's result reads as follows.

Theorem 11 (Stanley [24]). *The risk polynomial* $\mathcal{R}(\mathcal{G}; \mathbf{f})$ *equals the determinant of the* $(m-2) \times (m-2)$*-matrix* $\mathbf{I} + \mathbf{F}' \cdot \mathbf{A}$.

Example 12. Let $\mathcal{G}$ be the genotype lattice in Figure 1. Then $m = 8$ and $\mathbf{I} + \mathbf{F}' \cdot \mathbf{A}$ is the 6×6-matrix

$$\begin{array}{c} \\ 1000 \\ 0100 \\ 1100 \\ 0101 \\ 1110 \\ 1101 \end{array}
\begin{array}{c} \begin{array}{cccccc} 1000 & 0100 & 1100 & 0101 & 1110 & 1101 \end{array} \\
\begin{pmatrix}
1+f_{1000} & f_{1000} & 0 & f_{1000} & 0 & 0 \\
f_{0100} & 1+f_{0100} & 0 & 0 & 0 & 0 \\
f_{1100} & f_{1100} & 1+f_{1100} & f_{1100} & 0 & 0 \\
f_{0101} & f_{0101} & f_{0101} & 1+f_{0101} & f_{0101} & 0 \\
f_{1110} & f_{1110} & f_{1110} & f_{1110} & 1+f_{1110} & f_{1110} \\
f_{1101} & f_{1101} & f_{1101} & f_{1101} & f_{1101} & 1+f_{1101}
\end{pmatrix} \end{array}.$$

The determinant of this matrix is the risk polynomial of Example 6.

The Hilbert series method. A more conceptual way of thinking about the risk polynomial is based on the following algebraic construction. The *Stanley-Reisner ideal* $I_{\mathcal{G}'}$ of $\mathcal{G}'$ is the ideal generated by all quadratic monomials $f_g \cdot f_h$ where g and h are genotypes that are incomparable, i.e., neither $g \subseteq h$ nor $h \subseteq g$ holds. The ambient polynomial ring $S = \mathbb{R}[\mathbf{f}]$ is generated by the unknowns f_g where $g \in \mathcal{G}'$. The *Hilbert series* of $I_{\mathcal{G}'}$ is the formal sum over all monomials $\mathbf{f}^u = \prod_{g \in \mathcal{G}'} f_g^{u_g}$ which are not in the ideal $I_{\mathcal{G}'}$. This is a formal generating function which can be written as a rational function of the following form

$$H(S/I_{\mathcal{G}'}; \mathbf{f}) \quad = \quad \frac{K_{\mathcal{G}}(\mathbf{f})}{\prod_{g \in \mathcal{G}'}(1 - f_g)}.$$

Here $K_{\mathcal{G}}(\mathbf{f})$ is a polynomial in the unknowns f_g with integer coefficients. The polynomial $K_{\mathcal{G}}(\mathbf{f})$ is known as the *K-polynomial* of the ideal $I_{\mathcal{G}'}$. We refer to [17] for an introduction to Stanley-Reisner ideals and their K-polynomials.

If $\mathcal{E}$ is a directed forest (and we identify $f_g = p_g$) then Proposition 10 and [3, Thm. 14.11] imply that the ideal $I_{\mathcal{G}'}$ is an initial monomial ideal of the conjunctive Bayesian network on $\mathcal{E}$. In a forthcoming paper we shall prove that this initial ideal property holds for all event posets (not just trees).

Example 13. Let $\mathcal{G}$ be the genotype lattice in Figure 1. Then

$$I_{\mathcal{G}'} \quad = \quad \langle\, f_{0101}f_{1110},\, f_{1101}f_{1110},\, f_{0101}f_{1100},\, f_{0101}f_{1000},\, f_{0100}f_{1000} \rangle$$

is indeed the initial monomial ideal of the conjunctive Bayesian network in Example 9. The K-polynomial $K_{\mathcal{G}}(\mathbf{f})$ equals

$$\begin{gathered}1 - f_{0101}f_{1110} - f_{1101}f_{1110} - f_{0101}f_{1100} - f_{0101}f_{1000} - f_{0100}f_{1000} \\ +f_{0100}f_{1000}f_{0101} + f_{1000}f_{0101}f_{1100} + f_{1000}f_{0101}f_{1110} + f_{0101}f_{1100}f_{1110} \\ +f_{0101}f_{1110}f_{1101} + f_{0100}f_{1000}f_{1110}f_{1101} \\ -f_{1000}f_{0101}f_{1100}f_{1110} - f_{0100}f_{1000}f_{0101}f_{1110}f_{1101}.\end{gathered}$$

Again using Proposition 10 and Theorem 14.11 in [3] we see that the risk polynomial $\mathcal{R}(\mathcal{G};\mathbf{f})$ is the sum of all squarefree monomials in the expansion of the Hilbert series $H(S/I_{\mathcal{G}'};\mathbf{f})$. Equivalently, $\mathcal{R}(\mathcal{G};\mathbf{f})$ is the reduction of $H(S/I_{\mathcal{G}'};\mathbf{f})$ modulo the ideal generated by the squares f_g^2 of the unknowns. Since $1/(1-f_g)$ equals $1+f_g$ modulo $\langle f_g^2 \rangle$, we have the following result.

Proposition 14. *The risk polynomial $\mathcal{R}(\mathcal{G};\mathbf{f})$ of the genotype lattice $\mathcal{G}$ is the sum of all squarefree terms in the expansion of*

$$K_{\mathcal{G}}(\mathbf{f}) \cdot \prod_{g\in\mathcal{G}'} (1+f_g),$$

where $K_{\mathcal{G}}(\mathbf{f})$ is the K-polynomial of the Stanley-Reisner ideal $I_{\mathcal{G}'}$.

The univariate risk polynomial $\mathcal{R}(\mathcal{G};a)$ is derived from $\mathcal{R}(\mathcal{G};\mathbf{f})$ by replacing each f_g by the scalar unknown a. We have

$$\mathcal{R}(\mathcal{G};a) \quad = \quad c_0 + c_1 a + c_2 a^2 + \cdots + c_{n-1}a^{n-1},$$

where c_i is the number of chains of length i in $\mathcal{G}'$. Thus, $(c_0,\ldots,c_{n-1})$ is the f-vector of the simplicial complex of chains in $\mathcal{G}'$. Likewise, we get the graded risk polynomial from $\mathcal{R}(\mathcal{G};\mathbf{f})$ by replacing each f_g by $a_{|g|}$. We note that the graded risk polynomial is related to Ehrenborg's quasi-symmetric function encoding [12] of the flag f-vector of the chain complex of $\mathcal{G}'$.

The linear extensions method. One advantage of both Theorem 11 and Proposition 14 is that these formulas do not actually depend on the fact that $\mathcal{G}$ is a distributive lattice. They also apply if the set $\mathcal{G}$ of genotypes is an arbitrary poset. This is relevant for our discussion of the statistical models in Section 4, where we introduced a more general class of posets $\mathcal{G}_p \subseteq \{0,1\}^n$.

This advantage is also a disadvantage: Theorem 11 and Proposition 14 do not give the most efficient methods for computing $\mathcal{R}(\mathcal{G};\mathbf{f})$ when $\mathcal{G}$ is the distributive lattice induced by an event poset $\mathcal{E}$. In what follows we present a specialized and more efficient algorithm for the risk polynomial. The input to this algorithm consists of the event poset $\mathcal{E}$. It is not necessary to compute the genotype lattice $\mathcal{G}$ as this will be done as a byproduct of our approach, which is to compute the risk polynomial $\mathcal{R}(\mathcal{G};\mathbf{f})$ directly from $\mathcal{E}$.

As before, we assume that $\mathcal{E}$ has n elements, and we write $[n]$ for the linearly ordered set $\{1,2,\ldots,n\}$. A *linear extension* of $\mathcal{E}$ is an order-preserving bijection $\pi\colon \mathcal{E} \to [n]$. This means that $e < e'$ in $\mathcal{E}$ implies $\pi(e) < \pi(e')$. Every linear extension $\pi\colon \mathcal{E} \to [n]$ gives rise to an ordered list of $n-1$ genotypes $g^{(1)}, g^{(2)}, \ldots, g^{(n-1)}$ in $\mathcal{G}' = \mathcal{G}\backslash\{\hat{0}, \hat{1}\}$ as follows. The genotype $g^{(i)}$ is the subset of $\mathcal{E}$ consisting of all events whose image under π is among the first i positive integers. In symbols, $g^{(i)} = \pi^{-1}(\{1, 2, \ldots, i\})$. The sequence $g^{(1)}, g^{(2)}, \ldots, g^{(n-1)}$, derived from π, represents a mutational pathway in $\mathcal{G}$.

We now fix one distinguished linear extension of $\mathcal{E}$, that is, we identify the set underlying $\mathcal{E}$ with $[n]$ itself. Then a linear extension is simply any permutation π of $[n]$ which preserves the order relations in $\mathcal{E}$. We define

$$\mathbf{f}(\pi) \quad = \quad \prod_{i:\pi(i)<\pi(i+1)} (f_{g^{(i)}} + 1) \cdot \prod_{i:\pi(i)>\pi(i+1)} f_{g^{(i)}}, \tag{12}$$

where i runs over $\{1, 2, \ldots, n-1\}$. Our algorithm amounts to evaluating the risk polynomial by means of the following explicit summation formula.

Theorem 15. *The risk polynomial $\mathcal{R}(\mathcal{G}; \mathbf{f})$ equals the sum of the products $\mathbf{f}(\pi)$ where π runs over all linear extensions of the event poset $\mathcal{E}$.*

Proof. The relationship between chains in $\mathcal{G}$ and linear extensions of $\mathcal{E}$ is the content of [25, Prop. 3.5.2]. The distributive lattice $\mathcal{G}$ has a canonical *R-labeling* [25, Sec. 3.13] which assigns to each edge of the Hasse diagram of $\mathcal{G}$ the corresponding element of $\mathcal{E}$. In view of this R-labeling, Exercise 59d in [25, Chap. 3] tells us that the poset $\mathcal{G}' = \mathcal{G}\backslash\{\hat{0}, \hat{1}\}$ is *chain-partitionable.* Each product $\mathbf{f}(\pi)$ as in (12) is the generating function for all the chains in precisely one part of that chain partition of $\mathcal{G}'$. Adding up all products gives the generating function for all chains, which is the risk polynomial. □

Example 16. The event poset $\mathcal{E}$ in Figure 1 has five linear extensions π:

π	$\mathbf{f}(\pi)$
$(1,2,3,4)$	$(1+f_{1000})(1+f_{1100})(1+f_{1110})$
$(1,2,4,3)$	$(1+f_{1000})(1+f_{1100})f_{1101}$
$(2,1,3,4)$	$f_{0100}(1+f_{1100})(1+f_{1110})$
$(2,1,4,3)$	$f_{0100}(1+f_{1100})f_{1101}$
$(2,4,1,3)$	$(1+f_{0100})f_{0101}(1+f_{1101})$

The sum of these five products equals the risk polynomial $\mathcal{R}(\mathcal{G}; \mathbf{f})$.

Implementation. Pruesse and Ruskey [20] showed that the linear extensions of a poset $\mathcal{E}$ can be computed in time linear in the number of linear extensions. Thus, their algorithm computes $\mathcal{R}(\mathcal{G}; \mathbf{f})$ in time linear in the size of the output of Theorem 15. That output is in factored form (12) and is always more compact than the expanded risk polynomial. In this manner, we compute the risk polynomial in time sublinear in the size of the expanded risk polynomial.

To obtain the univariate risk polynomial, we take the sum of the terms $(1+a)^{n-1-\delta}a^{\delta}$, where $\delta = \delta(\pi)$ is the number of descents of the linear extension π. Similarly, the graded risk polynomial $\mathcal{R}(\mathcal{G}; a_1, \ldots, a_{n-1})$ is found by keeping track of the descent set of each linear extension π. We believe that this method is best possible for general posets $\mathcal{E}$. Notice that the leading term of the univariate risk polynomial is the number of linear extensions of $\mathcal{E}$, and it is #P-complete to count linear extensions [8].

When $\mathcal{E}$ is a directed forest, the recursive structure can be used to help compute the risk polynomial. In this case, $\mathcal{E}$ is built up by the operations of disjoint union and ordinal sum from the one element poset. For example, in the univariate case, the zeta polynomial [25, Sec. 3.11] of $\mathcal{G}$ behaves nicely under these operations and can be used to write down the risk polynomial. Based on these considerations, we can design an efficient algorithm for computing the univariate risk polynomial of a directed forest.

Using the method of Theorem 15, we have developed software for computing risk polynomials. The input to our program is an arbitrary event poset $\mathcal{E}$, and the output is the risk polynomial, the graded risk polynomial or the univariate risk polynomial. Optionally, the user can also input either exact fitness values or upper and lower bounds for each fitness value. The output in this case is either the exact risk of escape or upper and lower bounds for the risk. It is designed to integrate with the package `Mtreemix` [5], allowing the user to start with data, infer a mutagenetic tree, and then easily compute the risk polynomial. Our software is available at

`http://bio.math.berkeley.edu/riskpoly/`

We use the algorithm of [27] for computing linear extensions. Although this algorithm isn't asymptotically optimal, as shown in [20], it is simple to implement and efficient in practice.